BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 2000057152 A

(43) Date of publication of application: 25.02.00

(51) Int. Cl G06F 17/30

(21) Application number: 10222934

(22) Date of filing: 06.08.98

(71) Applicant: FUJI XEROX CO LTD

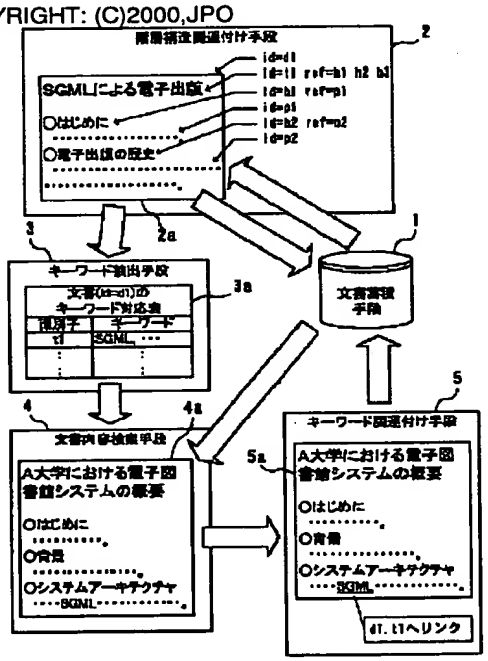
(72) Inventor: NUMATA KENICHI

(54) **DOCUMENT CORRELATING DEVICE, DOCUMENT ACCESSING DEVICE, COMPUTER-READABLE RECORDING MEDIUM RECORDING DOCUMENT CORRELATING PROGRAM AND COMPUTER-READABLE RECORDING MEDIUM RECORDING DOCUMENT READING PROGRAM**

(57) Abstract:

PROBLEM TO BE SOLVED: To make fast performable a processing which correlates a keyword in a document with a minimum related description in another document.

SOLUTION: A hierarchical structure correlating means 2 correlates the upper structure and lower structure of each element constituting a correlating object document 2a which is read from a document storing means 1. A keyword extracting means 3 extracts a keyword from a processing object element which has a specified attribute in the document 2a. A document content retrieving means 4 retrieves a document in the means 1 based on the extracted keyword. A keyword correlating means 5 correlates a keyword in the content of an extracted document 4a with the processing object element of the document 2a that becomes the extraction source of a keyword. It is possible to refer to related necessity minimum content in other documents by taking correlated information from a keyword in a document if such processing is carried out.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－57152
（P2000－57152A）
(43)公開日　平成12年２月25日(2000.2.25)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/40	３７０Ａ	５Ｂ０７５
			３４０	
		15/401	３１０Ａ	

審査請求　未請求　請求項の数９　ＯＬ　（全 19 頁）

(21)出願番号　特願平10－222934

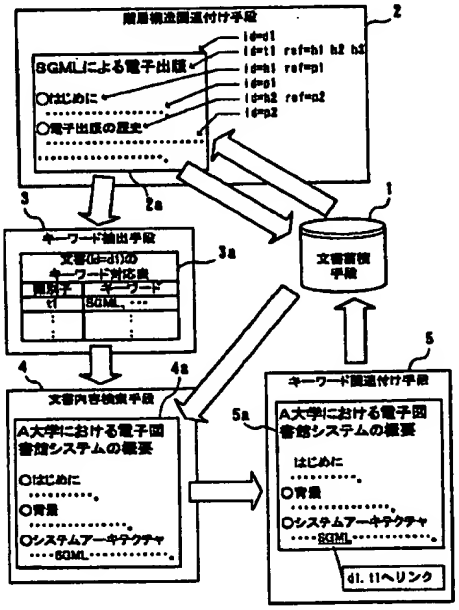
(22)出願日　平成10年８月６日(1998.8.6)

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号
(72)発明者　招田　賢一
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内
(74)代理人　100092152
弁理士　服部　毅巖

Ｆターム(参考)　5B075 ND03 ND35 ND38 NK02 NK32 NK43 NK44 NR03

(54)【発明の名称】　文書関連付け装置、文書閲覧装置、文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体及び文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体

(57)【要約】
【課題】　文書中のキーワードを他の文書中の最小限の関連記述に関連付ける処理を高速に行うことができるようにする。
【解決手段】　階層構造関連付け手段２は、文書蓄積手段１から読み込んだ被関連付け対象文書２ａを構成する各要素の上位構造と下位構造とを関連付ける。キーワード抽出手段３は、被関連付け対象文書２ａ中の特定の属性を有する処理対象要素からキーワードを抽出する。文書内容検索手段４は、抽出されたキーワードに基づいて、文書蓄積手段１中の文書を検索する。キーワード関連付け手段５は、検出された文書４ａの内容中のキーワードと、キーワードの抽出元となる被関連付け対象文書２ａの処理対象要素とを関連付ける。このような処理を実行すれば、文書中のキーワードから関連付けられた情報を取ることで、他の文書中で関連する必要最小限の内容を参照することができる。



1

【特許請求の範囲】

【請求項１】　文書間の関連付けを行う文書関連付け装置において、

階層的な論理構造の文書群を格納する文書蓄積手段と、

前記文書蓄積手段に格納されている文書を被関連付け対象文書とし、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段と、

前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードを抽出するキーワード抽出手段と、

前記キーワード抽出手段により抽出された前記キーワードを含む文書を、前記文書蓄積手段内より検索する文書内容検索手段と、

前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段と、

を有することを特徴とする文書関連付け装置。

【請求項２】　前記キーワード抽出手段は、前記被関連付け対象文書の表題としての属性を有する要素と、記載内容の見出しとしての属性を有する要素とを、前記処理対象要素として取り扱うことを特徴とする請求項１記載の文書関連付け装置。

【請求項３】　構造化文書の内容を閲覧する文書閲覧装置において、

階層的な論理構造の文書群を格納する文書蓄積手段と、

前記文書蓄積手段に格納されている被関連付け対象文書に対して、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段と、

前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容から、キーワードを抽出するキーワード抽出手段と、

前記キーワード抽出手段により抽出された前記キーワードに基づいて、前記文書蓄積手段に蓄積されている他の文書の内容を検索する文書内容検索手段と、

前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段と、

文書閲覧要求に応じて、前記文書蓄積手段から文書を抽出する文書抽出手段と、

前記文書抽出手段にて抽出された文書中で、前記キーワード関連付け手段により関連付けられた前記キーワードが選択されると、前記キーワードに対して関連付けられた前記被関連付け対象文書中の関連要素及び前記関連要素に関連付けられている下位の要素を順次抽出する要素抽出手段と、

前記要素抽出手段により抽出された前記関連要素の内容及び前記関連要素に関連付けられている下位の要素の内容を抽出する内容抽出手段と、

を有することを特徴とする文書閲覧装置。

2

【請求項４】　前記要素抽出手段により複数の前記関連要素が抽出された場合には、そのうちの１つの前記関連要素を選択し、選択した前記関連要素に関連付けられた下位の要素が複数存在する場合にはそのうちの１つの要素を選択する要素選択手段をさらに有し、

前記内容抽出手段は、前記要素選択手段により選択された前記関連要素の内容及び選択された下位の要素の内容を抽出する、

ことを特徴とする請求項３記載の文書閲覧装置。

【請求項５】　前記要素抽出手段は、前記キーワード関連付け手段により関連付けられた要素が複数抽出され、かつそれらの要素が同一文書内に存在する場合には、同一文書内に存在する要素への関連付けをグループ化して抽出することを特徴とする請求項３記載の文書閲覧装置。

【請求項６】　前記要素抽出手段は、文書ごとの関連付けをグループ化した場合には、同一文書内への関連付けの数、および関連付けられる要素の階層の深さから算出される重要度に応じて、各グループを並べ替えることを特徴とする請求項５記載の文書閲覧装置。

【請求項７】　前記要素抽出手段は、文書ごとにグループ化された関連付け要素群を、関連付けられる要素の階層の深さから算出される重要度および文書中での出現順序に応じてグループ内で並べ替えることを特徴とする請求項５記載の文書閲覧装置。

【請求項８】　文書間の関連付けを行うための文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体において、

階層的な論理構造の文書群を格納する文書蓄積手段、

前記文書蓄積手段に格納されている文書を被関連付け対象文書とし、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段、

前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードを抽出するキーワード抽出手段、

前記キーワード抽出手段により抽出された前記キーワードを含む文書を、前記文書蓄積手段内より検索する文書内容検索手段、

前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段、

としてコンピュータを機能させることを特徴とする文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項９】　構造化文書の内容を閲覧するための文書

閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体において、
階層的な論理構造の文書群を格納する文書蓄積手段、
前記文書蓄積手段に格納されている被関連付け対象文書に対して、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段、
前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容から、キーワードを抽出するキーワード抽出手段、
前記キーワード抽出手段により抽出された前記キーワードに基づいて、前記文書蓄積手段に蓄積されている他の文書の内容を検索する文書内容検索手段と、
前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段、
文書閲覧要求に応じて、前記文書蓄積手段から文書を抽出する文書抽出手段、
前記文書抽出手段にて抽出された文書中で、前記キーワード関連付け手段により関連付けられた前記キーワードが選択されると、前記キーワードに対して関連付けられた前記被関連付け対象文書中の関連要素及び前記関連要素に関連付けられている下位の要素を順次抽出する要素抽出手段、
前記要素抽出手段により抽出された前記関連要素の内容及び前記関連要素に関連付けられている下位の要素の内容を抽出する内容抽出手段、
としてコンピュータを機能させることを特徴とする文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は文書関連付け装置、文書閲覧装置、文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体、及び文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体に関し、特に文書中のあるキーワードとそのキーワードに関連する他の文書の内容を関連付ける文書関連付け装置、文書中のあるキーワードとそのキーワードに関連する他の文書の内容とが関連付けられた文書群中の文書を閲覧する文書閲覧装置、前記文書関連付け装置をコンピュータ上で実現するための文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体、及び前記文書閲覧装置をコンピュータ上で実現するための文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【０００２】

【従来の技術】ネットワーク上に散在する電子文書群をリンクによって関連付けることが可能な、いわゆるハイパーテキストシステムが、World Wide Web(WWW) の普及により、一般に広く利用されるようになってきている。ハイパーテキストシステムでは、ある文書中のキーワードに対して、より詳しい情報を持つ他の文書の内容へのハイパーリンクを付与しておく。これによって、利用者がその文書を閲覧していて、ハイパーリンクが付与された記述に関してより詳しく知りたいと思ったときには、そのハイパーリンクを辿ることによって関連情報を知ることができる。

【０００３】ところが、一般的にこのようなハイパーテキスト文書を作成するためには、文書の作成者が手作業でキーワードと他の文書との関連付けを行ってハイパーリンクを作成する必要があり、多大の労力と時間を要する。そこで、この問題を解決するために、文書中のキーワードを自動抽出して、他の文書から同一または同義のキーワードを含むものを検索することによって、文書間の関連付けすなわちハイパーリンクを自動的に作成することが考えられている。

【０００４】このとき単純に同一または同義のキーワードを手がかりとして文書を関連付けるだけでは、ハイパーリンクを辿ることによって、より詳しい説明が得られるという保証がない。なぜならば、関連付けられた文書のいずれにおいても同一または同義のキーワードが一言参照されているだけでそのキーワードの説明に当たる記述がない場合が往々にしてあり得るからである。

【０００５】この問題を解決する１つの方法として、特開平５－２０３６２号公報に開示された「文書テキスト間の連鎖自動作成システム」がある。この公報に開示された方法では、まず、文書テキストから重要キーワードを抽出し、抽出したキーワードの文書における重要度を算出する。その上で、同一のキーワードを共有する文書どうしで、キーワードの重要度の低い方の文書からキーワードの重要度の高い方の文書への、単方向の関連付けを自動生成する。この方法では、同一のキーワードを手がかりとして文書を関連付けているが、同一キーワードの文書における重要度の高い文書のほうが、重要度の低い文書よりも、そのキーワードに関してより詳しく説明されているものと仮定している。これによって、文書中のあるキーワードから、より詳しい説明が記述された他の文書に対するハイパーリンクが自動的に生成される。以下のこの方法を第１の従来技術とする。

【０００６】また、上記問題を解決する別の方法として特開平７－３２５８２７号公報に開示された「ハイパーテキスト自動生成装置」がある。この公報には、同一または同義のキーワードを持つ文書どうしを関連付ける際に、一方の文書のキーワードから、他の文書の同一または同義のキーワードを持つ章や節の見出しに対してハイパーリンクを生成する方法が示されている。この方法では、あるキーワードが見出しに含まれる場合、見出し以下の内容において、そのキーワードについて詳しく説明

されている可能性が高いと仮定している。これによって、文書中のあるキーワードから、より詳しい説明に対するハイパーリンクが自動的に生成される。以下のこの方法を第２の従来技術とする。

【０００７】

【発明が解決しようとする課題】しかし、いずれの従来技術においても、以下のような問題点があった。第１の従来技術では、関連付けの対象はある文書中のキーワードと他の文書全体である。そのため、関連付けられる他の文書の記述量が多い場合には、たとえ関連付けられたキーワードに対する詳しい説明が文書中に記述されていたとしても、文書中で関連する記述を見つけ出すことが困難である。

【０００８】第２の従来技術では、ある文書中のキーワードに対して、同一または同義のキーワードが含まれる他の文書が複数存在する場合には、予め与えられた戦略に従って候補をいずれか１つに絞るようになっている。そのため、利用者が実際に知りたい情報が関連付けの対象から洩れてしまうおそれがある。なお、この問題については、例えば関連付けの対象となる候補が複数存在する場合にその候補全てを関連付けてしまうことによって洩れを防ぐことができる。しかし、この場合には、利用者が複数の関連付けられた記述を順次閲覧し、必要な情報を探すという手間がかかる。

【０００９】さらに、上記２つの従来技術のいずれにおいても、関連付けの対象となるキーワードを自動抽出するために、文書全体に対して形態素解析を行う必要がある。形態素解析を高精度に行うには、かなり複雑な処理を行わなければならない。そのため、従来の技術を用いて大量の文書間のハイパーリンクを自動作成するには、処理に非常に時間がかかってしまうという問題点があった。

【００１０】本発明はこのような点に鑑みてなされたものであり、文書中のキーワードを他の文書中の最小限の関連記述に関連付ける処理を高速に行うことができる文書関連付け装置を提供することを目的とする。

【００１１】また、本発明の第２の目的は、文書中のキーワードを他の文書中の最小限の関連記述に関連付けられた文書群内の文書を閲覧するための文書閲覧装置を提供することである。

【００１２】また、本発明の第３の目的は、文書中のキーワードを他の文書中の最小限の関連記述に関連付ける処理をコンピュータに高速に行わせることができる文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することである。

【００１３】また、本発明の第４の目的は、文書中のキーワードを他の文書中の最小限の関連記述に関連付けられた文書群内の文書をコンピュータを用いて閲覧するための文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体を提供することである。

【００１４】

【課題を解決するための手段】本発明では上記課題を解決するために、文書間の関連付けを行う文書関連付け装置において、階層的な論理構造の文書群を格納する文書蓄積手段と、前記文書蓄積手段に格納されている文書を被関連付け対象文書とし、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段と、前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードを抽出するキーワード抽出手段と、前記キーワード抽出手段により抽出された前記キーワードを含む文書を、前記文書蓄積手段内より検索する文書内容検索手段と、前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段と、を有することを特徴とする文書関連付け装置が提供される。

【００１５】このような文書関連付け装置によれば、階層構造関連付け手段により、前記文書蓄積手段に格納されている文書が被関連付け対象文書とされ、その被関連付け対象文書を構成する各要素の上位構造と下位構造とが関連付けられる。また、キーワード抽出手段により、被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードが抽出される。すると、内容検索手段により、キーワード抽出手段が抽出したキーワードを含む文書が文書蓄積手段内から検索される。そして、キーワード関連付け手段により、文書内容検索手段により検出された文書中のキーワードと、キーワードの抽出元となる被関連付け対象文書内の処理対象要素とが関連付けられる。

【００１６】また上記課題を解決するために、構造化文書の内容を閲覧する文書閲覧装置において、階層的な論理構造の文書群を格納する文書蓄積手段と、前記文書蓄積手段に格納されている被関連付け対象文書に対して、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段と、前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容から、キーワードを抽出するキーワード抽出手段と、前記キーワード抽出手段により抽出された前記キーワードに基づいて、前記文書蓄積手段に蓄積されている他の文書の内容を検索する文書内容検索手段と、前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段と、文書閲覧要求に応じて、前記文書蓄積手段から文書を抽出する文書抽出手段と、前記文書抽出手段にて抽出された文書中で、前記キーワード関連付け手段により関連付けられた前記キーワードが選択されると、前記キーワードに対して関連付けられた前記被関連付け対象文書中の関連要素及び前記関連要素

に関連付けられている下位の要素を順次抽出する要素抽出手段と、前記要素抽出手段により抽出された前記関連要素の内容及び前記関連要素に関連付けられている下位の要素の内容を抽出する内容抽出手段と、を有することを特徴とする文書閲覧装置が提供される。

【００１７】このような文書閲覧装置によれば、階層構造関連付け手段により、前記文書蓄積手段に格納されている文書が被関連付け対象文書とされ、その被関連付け対象文書を構成する各要素の上位構造と下位構造とが関連付けられる。また、キーワード抽出手段により、被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードが抽出される。すると、内容検索手段により、キーワード抽出手段が抽出したキーワードを含む文書が文書蓄積手段内から検索される。そして、キーワード関連付け手段により、文書内容検索手段により検出された文書中のキーワードと、キーワードの抽出元となる被関連付け対象文書内の処理対象要素とが関連付けられる。さらに、文書閲覧要求が入力されると、文書抽出手段により、文書閲覧要求に応じた文書が文書蓄積手段から抽出される。この文書抽出手段にて抽出された文書中で、キーワード関連付け手段により関連付けられたキーワードが選択されると、要素抽出手段により、キーワードに対して関連付けられた被関連付け対象文書中の関連要素及び関連要素に関連付けられている下位の要素が順次抽出される。

【００１８】さらに、内容抽出手段により、前記要素抽出手段により抽出された前記関連要素の内容及び関連要素に関連付けられている下位の要素の内容が抽出される。また上記課題を解決するために、文書間の関連付けを行うための文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体において、階層的な論理構造の文書群を格納する文書蓄積手段、前記文書蓄積手段に格納されている文書を被関連付け対象文書とし、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段、前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容からキーワードを抽出するキーワード抽出手段、前記キーワード抽出手段により抽出された前記キーワードを含む文書を、前記文書蓄積手段内より検索する文書内容検索手段、前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段、としてコンピュータを機能させることを特徴とする文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体が提供される。

【００１９】この記録媒体に記録された文書関連付けプログラムをコンピュータに実行させれば、上記本発明に係る文書関連付け装置の機能がコンピュータ上に構築される。

【００２０】また上記課題を解決するために、構造化文書の内容を閲覧するための文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体において、階層的な論理構造の文書群を格納する文書蓄積手段、前記文書蓄積手段に格納されている被関連付け対象文書に対して、前記被関連付け対象文書を構成する各要素の上位構造と下位構造とを関連付ける階層構造関連付け手段、前記被関連付け対象文書中の特定の属性を有する処理対象要素に含まれる内容から、キーワードを抽出するキーワード抽出手段、前記キーワード抽出手段により抽出された前記キーワードに基づいて、前記文書蓄積手段に蓄積されている他の文書の内容を検索する文書内容検索手段と、前記文書内容検索手段により検出された文書中の前記キーワードと、前記キーワードの抽出元となる前記被関連付け対象文書内の前記処理対象要素とを関連付けるキーワード関連付け手段、文書閲覧要求に応じて、前記文書蓄積手段から文書を抽出する文書抽出手段、前記文書抽出手段にて抽出された文書中で、前記キーワード関連付け手段により関連付けられた前記キーワードが選択されると、前記キーワードに対して関連付けられた前記被関連付け対象文書中の関連要素及び前記関連要素に関連付けられている下位の要素を順次抽出する要素抽出手段、前記要素抽出手段により抽出された前記関連要素の内容及び前記関連要素に関連付けられている下位の要素の内容を抽出する内容抽出手段、としてコンピュータを機能させることを特徴とする文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体が提供される。

【００２１】この記録媒体に記録された文書閲覧プログラムをコンピュータに実行させれば、上記本発明に係る文書閲覧装置の機能がコンピュータ上に構築される。

【００２２】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。図１は、本発明の原理構成図である。本発明の文書関連付け装置は、以下の要素で構成される。

【００２３】文書蓄積手段１は、階層的な論理構造の文書群を蓄積する。構造化された文書としては、ＳＧＭＬの規定に従って作成された文書などがある。階層構造関連付け手段２は、文書蓄積手段１から被関連付け対象文書２ａを読み込み、読み込んだ被関連付け対象文書２ａを構成する各要素の上位構造と下位構造とを関連付ける。例えば、各要素に識別子を与える。そして、各要素に対して、その要素の下位構造となる要素の識別子の情報を持たせる。要素間の関連付けを行った被関連付け対象文書２ａは、文書蓄積手段１に戻す。

【００２４】キーワード抽出手段３は、被関連付け対象文書２ａ中の特定の属性を有する処理対象要素に含まれる内容からキーワードを抽出する。例えば、表題としての属性を有する要素と、見出しとしての属性を有する要

素とを、処理対象要素とする。すると、キーワード抽出手段３は、抽出元の要素の識別子と、その要素から抽出されたキーワードの集合とを対応づけたキーワード対応表３ａを内部で生成する。そして、被関連付け対象文書２ａに関するキーワード対応表３ａを文書内容検索手段４に渡す。

【００２５】文書内容検索手段４は、キーワード抽出手段３により抽出されたキーワードに基づいて、文書蓄積手段１に蓄積されている他の文書の内容を検索する。見つけ出した文書４ａは、キーワード関連付け手段５に渡す。

【００２６】キーワード関連付け手段５は、文書内容検索手段４により検出された文書４ａの内容中のキーワードと、キーワードの抽出元となる被関連付け対象文書２ａの処理対象要素とを関連付ける。被関連付け対象文書２ａの特定の要素への関連付けを行った文書５ａは、文書蓄積手段１に格納する。

【００２７】このような文書関連付け装置によれば、階層構造関連付け手段２に読み込まれた被関連付け対象文書２ａは、各要素の上位構造と下位構造との関連付けが行われ、文書蓄積手段１に戻される。このとき、キーワード抽出手段３により、各要素の内容の中からキーワードが抽出される。すると、文書内容検索手段４により、抽出されたキーワードに基づいて文書蓄積手段１内の文書が検索される。検出された文書４ａはキーワード関連付け手段５に渡され、文書４ａの内容中のキーワードと、キーワードの抽出元となる被関連付け対象文書２ａの処理対象要素とが関連付けられる。そして、処理対象要素との関連付けが行われた文書５ａは、文書蓄積手段１に戻される。

【００２８】このような処理を、文書蓄積手段１に格納されている全ての文書を被関連付け対象文書２ａとして実行すれば、ある文書中のキーワードが他の文書中の特定の要素（表題や見出し）に関連付けられ、さらに、その要素から下位構造に関連付けられる。そのため、文書蓄積手段１内の文書を閲覧する場合には、文書中のキーワードから他の文書中の必要最小限の関連付けられた内容を参照することができる。

【００２９】しかも、関連付けに際して文書中の表題もしくは見出しなどの特定の要素だけを対象としてキーワード抽出処理を行うので、形態素解析のようなキーワード抽出に必要な煩雑な処理を文書全体に対して施す必要がなくなる。その結果、関連付けの処理効率が向上する。

【００３０】次に、本発明の文書関連付け装置によって文書間の関連付けを行い、それらの文書を閲覧することができる文書閲覧装置を第１の実施の形態として以下に説明する。

【００３１】図２は、本発明を適用した文書閲覧装置の構成を示す図である。この文書閲覧装置は、文書蓄積部１１、階層構造関連付け部１２、キーワード抽出部１３、文書内容検索部１４、キーワード関連付け部１５、文書抽出部１６、見出し抽出部１７、見出し選択部１８、内容抽出部１９、表示部２０、及び入力部２１から構成されている。

【００３２】文書蓄積部１１は、表題、章の見出し、節の見出し、段落等の論理構造を有する文書群を蓄積する。階層構造関連付け部１２は、文書蓄積部１１に蓄積された文書を読み込み、表題、見出しの階層（章見出し、節見出しなど）、見出しに対応する内容（例えばある節の段落の並び）を関連付ける。

【００３３】キーワード抽出部１３は、階層構造関連付け部１２にて関連付けられた表題および見出しの階層からキーワードを抽出する。文書内容検索部１４は、キーワード抽出部１３にて抽出されたキーワードを用いて、文書蓄積部１１に蓄積された文書群を対象に、与えられたキーワードを内容に持つ文書を検索する。

【００３４】キーワード関連付け部１５は、文書内容検索部１４にて検索された文書中のキーワードと、該キーワードを抽出した表題および見出しの階層を関連付ける。文書抽出部１６は、文書蓄積部１１に蓄積された文書群から、入力部２１で利用者が入力した要求に応じて文書を抽出し、表示部２０に表示する。

【００３５】見出し抽出部１７は、文書抽出部１６により抽出され、表示部２０に表示された文書中で、利用者が入力部２１によりキーワードを指定した場合に、指定されたキーワードで関連付けられている他の文書の表題もしくは見出しを文書蓄積部１１から抽出し、表示部２０に表示する。また、抽出された前記表題もしくは見出しのさらに下位の見出しを文書蓄積部１１から抽出し、表示部２０に表示する。

【００３６】見出し選択部１８は、入力部２１で利用者が入力した要求に応じて、見出し抽出部１７により表題もしくは見出しが複数抽出された場合にはそのうちの１つの表題もしくは見出しを選択し、前記表題もしくは見出しに下位の見出しが複数存在する場合にはそのうちの１つの見出しを選択する。

【００３７】内容抽出部１９は、見出し抽出部１７により抽出された表題、見出しもしくは順次抽出された下位の見出しが、その見出しに対応する内容と関連付けられている場合に、文書蓄積部１１からその内容を抽出し、表示部２０に表示する。

【００３８】表示部２０は、文書抽出部１６により抽出された文書、見出し抽出部１７により抽出された他の文書の表題もしくは見出し、および内容抽出部１９により抽出された他の文書の内容を、画面上に表示する。

【００３９】入力部２１は、文書抽出部１６により抽出する文書の指定、文書抽出部１６により抽出された文書中でのキーワードの選択、見出し抽出部１７により抽出された表題もしくは見出しが複数存在する場合の選択の

指示等を行う。

【００４０】次に、このような構成の文書閲覧装置により、文書蓄積部１１に格納されている文書群に対して文書間の関連付けを行う手順について説明する。図３は、文書間の関連付けを行う手順を示すフローチャートである。以下の処理をステップ番号に沿って説明する。

［Ｓ１］階層構造関連付け部１２が、文書蓄積部１１から未処理の文書を１つ読み込む。

［Ｓ２］階層構造関連付け部１２が、読み込んだ文書の構造を解析する。

［Ｓ３］階層構造関連付け部１２が、表題、見出し、及び内容を関連付ける。

［Ｓ４］キーワード抽出部１３が、表題及び見出しの内容の中からキーワードを抽出する。

［Ｓ５］文書内容検索部１４が、キーワード抽出部１３が抽出したキーワードを含む文書を、文書蓄積部１１の中から検索する。

［Ｓ６］キーワード関連付け部１５が、文書内容検索部１４によって検出された文書内のキーワードに合致した部分に対して、そのキーワードの抽出元となった表題もしくは見出しを関連付ける。

［Ｓ７］キーワード関連付け部１５が、キーワードの関連付けの終了した文書を文書蓄積部１１へ格納する。

［Ｓ８］階層構造関連付け部１２は、文書蓄積部１１に格納されている全ての文書の処理を行ったか否かを判断し、全ての文書に対する処理が終了していれば文書間の関連付け処理を終了し、そうでなければステップＳ１に進み未処理の文書に対する処理を行う。

【００４１】このような処理を行うことにより、各文書の内容に含まれるキーワードから、そのキーワードを表題もしくは見出しとして含む文書の該当する表題若しくは見出しへリンクを張ることができる。

【００４２】以下に、具体例を用いて処理内容の詳細を説明する。なお、以下の例では、表題、見出し等の論理構造を有する文書の一例として、国際規格であるＳＧＭＬ(Standard Generalized Markup Language; ISO8879)に基づく表現を用いているが、表題、見出し、見出しに対応する内容が表現できる体系であればＳＧＭＬでなくともよい。

【００４３】まず、階層構造関連付け部１２が、文書蓄積部１１に蓄積された文書を１つ読み込む（ステップＳ１）。ここで、以下のような文書を読み込んだものとする。図４は、関連付けの対象となるキーワードを見出しに含む文書の第１の例を示す図である。この文書３１は、以下のような構造定義に従って作成されている。

【００４４】文書中の各要素は、その開始と終了を示すタグによって囲まれている。ある要素Ａについて、開始タグは＜Ａ＞、終了タグは＜／Ａ＞で示される。文書は、文書の開始を示すタグ＜doc ＞と、文書の終了を示すタグ＜／doc ＞によって囲まれている。文書要素(doc)は表題を示す要素(title)と章を示す要素(sect1)の並びとを包含している。章要素(sect1)は見出しを示す要素(head)と段落を示す要素(para)の並びとを包含しているか、もしくは、見出し要素(head)と節を示す要素(sect2)の並びを包含している。節要素(sect2)は見出し要素(head)と段落要素(para)の並びを包含している。また、表題要素(title)、見出し要素(head)、段落要素(para)は、その内容としてテキスト（文字列）を持つ。

【００４５】なお、本実施の形態で例示する文書では、要素の名前としてdoc、title、sect1、sect2、head、paraを用いているが、文書中で表題、見出し、本文が特定できれば、名前はなんでもよい。また、章や節の構造はさらに深く入れ子になっていてもよい。例えば、節要素(sect2)がさらに下位の節要素(sect3)を含むようになっていてもよい。

【００４６】このような文書３１を読み込んだ階層構造関連付け部１２は、読み込んだ文書の表題、見出し、段落等の文書構造を解析し、文書中の各要素に一意な識別子を付与する（ステップＳ２）。

【００４７】図５は、各要素に一意な識別子を付与した文書を示す図である。この図では、各要素に属性名「id」の値として識別子を付与している。この文書３２では、文書要素(doc)に「d1」という識別子を付与している。文書要素の識別子が、文書３２自身の識別子となる。そのため、文書要素の識別子は、文書蓄積部１１に格納されている文書の中で一意に識別できるような記号が用いられる。

【００４８】文書３２中の文書要素以外の要素に関しては、文書３２内において一意に識別できればよい。ここでは、表題要素(title)に「t1」という識別子を付与し、章要素(sect1)にそれぞれ「s1」、「s2」、「s3」という識別子を付与し、見出し要素(head)にそれぞれ「h1」、「h2」、「h3」という識別子を付与し、段落要素(para)にそれぞれ「p1」、「p2」、「p3」、「p4」という識別子を付与している。

【００４９】次に、階層構造関連付け部１２は文書３２の表題、見出し、もしあれば下位の見出し、見出しに対応する段落の並びを関連付ける（ステップＳ３）。本実施の形態では、文書の表題から見出しへの関連付けを、表題要素(title)の属性として見出しの識別子の並びを設定することによって表現する。また、見出しから下位の見出しへの関連付けもしくは見出しから対応する内容への関連付けは、見出し要素(head)の属性として下位の見出し要素の識別子もしくは内容となる段落要素(para)の識別子の並びを設定することによって表現する。

【００５０】図６は、表題、見出し、内容を関連付けた文書の例を示す図である。この文書３３は、図５に示す文書３２の表題要素および見出し要素に、関連付ける見出し要素もしくは段落要素の識別子の並びを属性名「ref」の値として付与したものである。この例では、識別

子の並びを空白文字によって区切っている。例えば、表題要素(title) の下位には３つの見出し要素(head)があるため、表題要素(title) の属性名「ref 」の値は、「h1 h2 h3」となる。

【００５１】次に、キーワード抽出部１３が階層構造関連付け部１２によって関連付けられた表題もしくは見出しからキーワードを抽出する（ステップＳ４）。キーワードの抽出方法としては、従来の形態素解析などの手法を利用すればよい。本実施の形態では、形態素解析の結果から名詞と判定された単語をキーワードとして利用する。また、ひらがな語など、キーワードになりにくいものは、予めストップワードとして登録しておき、キーワードの抽出対象から外す。キーワード抽出部１３は、要素と、その要素に含まれるキーワードとの対応関係を示すキーワード対応表を作成し、一時的に保持する。

【００５２】図７は、キーワード対応表の例を示す図である。これは、図６に示した文書３３の表題要素(title) および見出し要素(head)と、そこから抽出したキーワードとの対応関係を示すキーワード対応表４１である。キーワード対応表４１には、「要素の種類」、「識別子」、および「キーワード」の項目が設けられている。「要素の種類」の項目には、キーワードの抽出を行った要素の種類が設定される。この例は、「表題」か「見出し」のいずれかである。「識別子」の項目には、キーワードの抽出を行った要素の識別子が設定される。「キーワード」の項目には、キーワードの抽出を行った要素に含まれていたキーワードの集合が設定される。

【００５３】このように、文書中の表題要素および見出し要素のみに対して形態素解析処理を行うので、文書全体に対して形態素解析処理を行う必要はない。一般に文書の表題や見出しに含まれるテキストの量は、文書全体のテキスト量に比して非常に少ないので、形態素解析の処理コストを大幅に削減することができる。

【００５４】次に、文書内容検索部１４は、キーワード抽出部１３により抽出されたキーワードを用いて、文書蓄積部１１に蓄積された他の文書の内容を検索する（ステップＳ５）。例えば、表題要素(title) から抽出された「SGML」というキーワードを用いて、文書蓄積部１１内の文書を検索を行った場合、以下のような文書が検出される。

【００５５】図８は、関連付けの対象となるキーワードを本文中に含む文書の例を示す図である。この文書５１は、段落要素(para)の内容に含まれるテキスト「...SGML へ変換する。... 」の「SGML」が一致したことにより、検出される。なお、この文書５１は、図４に示した文書３１と同様の構造定義に従って作成された文書である。

【００５６】図８のような文書５１が見つかったら、そのキーワード関連付け部１５はキーワードと一致する文書５１の内容と、そのキーワードを含む表題もしくは見出しを関連付ける（ステップＳ６）。具体的には、テキスト「...SGML へ変換する。... 」中の「SGML」を参照元要素としてタグ付けし、図６に示した文書３３の表題要素(title) の識別子を、参照元の要素の属性として設定する。

【００５７】図９は、キーワードと表題との関連付けが行われた文書の例を示す図である。この文書５２では、キーワード「SGML」は関連付けを示す要素(link)の開始タグと終了タグによって囲まれ、link要素の属性「ref 」の値として文書「d1」の表題「t1」への関連付けが設定されている。ここで属性「ref 」の値として、文書要素の識別子「d1」と表題要素の識別子「t1」を「. 」によって接続しているのは、識別子「t1」が他の文書のある要素においてたまたま使われている場合に、関連付けの対象を一意に決定できなくなることを防ぐためである。

【００５８】なお、本実施の形態では文書要素の識別子と表題要素もしくは見出し要素とを接続するために「. 」を用いているので、要素に識別子を付与する際には識別子自身に「. 」を含めないようにする。

【００５９】また、本実施の形態では、文書要素(doc) の識別子が、文書蓄積部１１に蓄積されている文書を一意に識別できるように付与されているため、この文書要素を用いて文書を識別しているが、文書を識別するための識別子を文書全体に対して付与して、それを関連付けの識別子として用いてもよい。このような識別子としては、文書の実体がファイルである場合にはファイル名を用いたり、文書がWWW(World Wide Web)上で公開される場合にはＵＲＬ(Uniform Resource Locator)を用いたりすることができる。

【００６０】ステップＳ４にて抽出された全てのキーワードに対して他の文書内容を検索し、ステップＳ６にてキーワードの関連付けが終了したら、関連付けされた文書は文書蓄積部１１に格納される（ステップＳ７）。このとき、関連付けの対象となった元の文書の内容は上書きされる。

【００６１】そして、文書蓄積部１１に蓄積された全ての文書について、上記ステップＳ１～ステップＳ７の処理が行われたかどうかを調べ（ステップＳ８）、まだ処理されていない文書があればステップＳ１へ戻って処理を継続し、全ての文書について処理が終了していれば、文書間の関連付けの処理を終了する。

【００６２】以上の処理が行われることにより、図９に示した文書５２に対しても、階層構造の関連付けが行われる。図１０は、図９の文書に対して階層構造の関連付けを行った結果を示す図である。この文書５３は、文書要素(doc) の識別子として「d2」が付与されている。

【００６３】次に、本発明に基づく文書関連付け装置により、関連付けを利用して、文書中のあるキーワードから、そのキーワードに対する説明記述を参照する手順に

ついて説明する。

【００６４】図１１は、関連付けの利用手順を示すフローチャートである。このフローチャートをステップ番号に沿って簡単に説明する。

［Ｓ１１］利用者が入力部２１を用いて文書の表示要求を入力すると、文書抽出部１６が該当する文書を文書蓄積部１１内から抽出する。抽出した文書の内容は、表示部２０の画面に表示される。

［Ｓ１２］利用者が入力部２１を用いてキーワードを選択する。

［Ｓ１３］見出し抽出部１７が、ステップＳ１２にて選択されたキーワードの関連付け情報すなわちlink要素の属性「ref 」の識別子を参照し、文書蓄積部１１から該当する識別子を持つ文書の表題もしくは見出しを抽出する。あるいは後述するステップＳ１４，Ｓ１５で見出し選択部１８によって選択された表題もしくは見出しの下位の見出しを、文書蓄積部１１から抽出する。そして、抽出した表題もしくは見出しを表示部２０に表示する。

［Ｓ１４］見出し選択部１８が、見出し抽出部１７によって抽出された見出しが複数か否かを判断し、複数であればステップＳ１５へ処理を進め、１つだけであればその表題もしくは見出しを選択してステップＳ１６へ処理を進める。

［Ｓ１５］見出し選択部１８が、入力部２１で利用者が入力した要求に応じて、見出し抽出部１７により表題もしくは見出しが複数抽出された場合にはそのうちの１つの表題もしくは見出しを選択する。

［Ｓ１６］見出し選択部１８は、選択された表題もしくは見出しに関して、下位の見出しが存在するか否かを判断する。この実施の形態では、ステップＳ１３にて抽出された表題要素(title) もしくは見出し要素(head)の属性「ref 」の値として設定されている識別子を持つ要素を特定し、その要素が見出し要素(title)であるかないかを判定する。下位の見出しが存在していればステップＳ１３に進み、存在していなければステップＳ１７に進む。

［Ｓ１７］内容抽出部１９が、ステップＳ１５にて選択された見出し要素に関連付けられた内容に対応する要素を抽出し、表示部２０の画面に表示する。

【００６５】以下に、関連付けの利用に関する処理を具体例を用いて説明する。まず利用者が図１０に示した文書５３の表示要求を入力部２１により指示したものとする。すると、文書５３の内容が表示部２０の画面に表示される。

【００６６】図１２は、文書の内容を表示した際の表示画面の例を示す図である。この表示画面６１では、文書中のタグにより表題、見出し、段落、関連付けられたキーワードなどを識別し、それぞれに対して適切なレイアウトを定めて画面表示を行っている。例えば表題は大きめのフォントでセンタリングして表示し、見出しは大きめのフォントで番号を付与して表示し、他の文書の見出し等に関連付けられたキーワードは下線を付与して強調している。

【００６７】次に、利用者が、表示部２０に表示された文書を参照し、関連付けの付与された「SGML」の表示箇所をマウスでクリックするなどの方法で選択したものとする（ステップＳ１２）。すると、見出し抽出部１７が、選択されたキーワード「SGML」の関連付け情報すなわちlink要素の属性「ref 」の識別子を参照し、文書蓄積部１１から該当する識別子「d1」を持つ文書３３内の該当する表題「t1」を抽出し、表示部２０に表示する（ステップＳ１３）。

【００６８】図１３は、見出しを表示した際の表示画面の例を示す図である。前述の関連付けの処理によりキーワード「SGML」は関連付けを示すlink要素によってタグ付けされており、その属性「ref 」の値として「d1.t1 」が設定されているので、図６に示した文書３３の表題要素(識別子は「t1」) が見出し抽出部１７により抽出され、表題要素の内容「SGMLによる電子出版」を含む表示画面６２が、表示部２０により表示される。

【００６９】このとき、抽出された表題が複数か否かの判定が見出し抽出部１７によって行われるが（ステップＳ１４）、この例では抽出された表題もしくは見出しが１つだけである。そこで、見出し抽出部１７は、抽出された見出しに関連付けられた下位の見出しが存在するかどうかを判定する（ステップＳ１６）。この例では、識別子「t1」を持つ表題要素の属性「ref 」の値として、「h1 h2 h3」の３つの要素が関連付けられており、いずれも見出し要素である。従って、ステップＳ１３へ戻り見出しの抽出が行われる。

【００７０】図１４は、下位の見出しを表示した際の表示画面の例を示す図である。これは、図１３に示した表示画面６２の例から、「SGMLによる電子出版」を内容に持つ表題要素に関連付けられている下位の見出しを表示部２０に表示したときの表示画面６３の例を示したものである。すなわち、図６に示した文書３３において、識別子「t1」を持つ表題要素の属性「ref 」の値として設定されている３つの見出し要素(識別子はh1、h2、h3) の内容「はじめに」「電子出版の歴史」「関連ツール」を抽出し、表示部２０の画面に表示している。

【００７１】ここで、再び見出し選択部１８が、抽出された見出しが複数であるか否かの判断を行う（ステップＳ１４）。ここでは、３つの見出しが抽出されているので、利用者は表示部２０に表示されている複数の表題もしくは見出しから入力部２１により１つを選択する（ステップＳ１５）。この例では、図１４に表示されている３つの見出しの内容のうち「関連ツール」をマウス等で選択したものとする。

【００７２】すると、見出し選択部１８が、選択された見出し「関連ツール」に関連付けられた下位の見出しが

存在するかどうかを判定する（ステップＳ１６）。図６に示した文書３３において、「関連ツール」を内容に持つ見出し要素(識別子は「h3」）の属性「ref 」の値として設定されている識別子p3、p4、．．．の要素はいずれも見出しではない。したがって、内容抽出部１９が、内容の抽出を行う（ステップＳ１７）。

【００７３】図１５は、内容を表示した際の表示画面の例を示す図である。これは、図１４に示した表示画面６３の例から、「関連ツール」を内容に持つ見出し要素に関連付けられている内容を表示部２０に表示したときの表示画面６４の例である。すなわち、図６に示した文書３３において、識別子「h3」を持つ見出し要素の属性「ref 」の値として設定されている段落要素（識別子p3、p4、．．．）の内容を抽出し、表示部２０に表示する。

【００７４】このように、関連する内容の候補が複数存在する場合にも、見出しを表示して選択することにより必要最小限の関連付けられた内容を参照することができる。また、表示部２０に表示される表題もしくは見出しから、利用者が内容を参照する必要がないと判断した場合は、内容の参照を行う前に処理を中断することも可能である。したがって、利用者は内容の詳細を全て読むことなく必要な情報を効率良く見つけることが可能である。

【００７５】次に、第２の実施の形態について説明する。第２の実施の形態は、ある文書内容中のキーワードに対して、他の文書の表題もしくは見出しが複数関連付けられている場合に、関連付けられた内容をさらに効率的に抽出できるようにした文書閲覧装置である。なお、第２の実施の形態の構成要素は、図２に示した第１の実施の形態の構成要素と同じであるため、図２に示した構成を用いて第２の実施の形態を説明する。また、第２の実施の形態における文書間の関連付け処理は、第１の実施の形態と同様であるため説明を省略する。

【００７６】そこで、第２の実施の形態による関連付け参照処理について、以下に説明する。図１６は、第２の実施の形態における関連付け参照の処理の流れを示すフローチャートである。以下の処理をステップ番号に沿って説明する。

［Ｓ２１］利用者が文書蓄積部１１に蓄積された文書群から抽出する文書を入力部２１により指示すると、文書抽出部１６は、指示された文書を抽出し、表示部２０に表示する。

［Ｓ２２］利用者が表示部２０に表示された文書を参照し、入力部２１より関連付けの付与されたキーワードの表示箇所をマウスでクリックするなどの方法で選択する。

［Ｓ２３］見出し抽出部１７は、ステップＳ２２にて選択されたキーワードの関連付け情報すなわちlink要素の属性「ref 」の識別子を参照し、文書蓄積部１１から該当する識別子を持つ文書の表題もしくは見出しを抽出する。

［Ｓ２４］見出し抽出部１７は、ステップＳ２３にて抽出された表題もしくは見出しが１つであるか複数であるかを判定し、抽出された表題もしくは見出しが複数あれば、ステップＳ２５へ進み、１つしかなければステップＳ２９へ進む。

［Ｓ２５］見出し抽出部１７は、ステップＳ２４にて抽出された表題もしくは見出しが複数あると判定されると、それらの表題もしくは見出しを文書ごとにグループ化する。

［Ｓ２６］見出し抽出部１７は、ステップＳ２５にてまとめられた文書ごとの関連付けのグループを、同一文書内への関連付けの数、および関連付けられる表題もしくは見出しの階層の深さから算出される重要度に応じて並べ替える。

［Ｓ２７］見出し抽出部１７は、ステップＳ２５にて文書ごとにグループ化された関連付けを、関連付けられる表題もしくは見出しの階層の深さから算出される重要度に応じて各グループ内で並び替える。

［Ｓ２８］利用者は表示部２０に表示されている複数の表題もしくは見出しから入力部２１により１つを選択する。

［Ｓ２９］見出し抽出部１７は、ステップＳ２３にて抽出された表題もしくは見出しが１つである場合またはステップＳ２８にて見出しが選択された場合に、その表題もしくは見出しに関連付けられた下位の見出しが存在するかどうかを判定する。もし下位の見出しが存在すればステップＳ２３に戻って下位の見出しを抽出する。下位の見出しが存在しなければステップＳ３０へ進む。

［Ｓ３０］内容抽出部１９が、ステップＳ２８にて選択された見出し要素に関連付けられた内容に対応する要素を抽出し、表示部２０の画面に表示する。

【００７７】このようにして、ある文書内容中のキーワードに対して、他の文書の表題もしくは見出しが複数関連付けられている場合に、関連付けられた内容を効率的に抽出することができる。以下にこの処理の詳細を、具体例を用いて説明する。

【００７８】本実施の形態では、第１の実施の形態で示した文書以外に、関連付けの対象となるキーワード「SGML」を表題に含む次のような文書が、文書蓄積部１１に格納されているものとする。

【００７９】図１７は、関連付けの対象となるキーワードを表題に含む文書の第２の例を示す図である。この文書７１には、文書要素(doc) に「d3」という識別子が付与されている。また、「id="t1" 」の表題要素(title)、「id="h2" 」の見出し要素(head)、および「id="h3" 」の見出し要素(head)の内容に「SGML」のキーワードが含まれている。

【００８０】図１８は、関連付けの対象となるキーワー

ドを表題に含む文書の第３の例を示す図である。この文書８１には、文書要素(doc) に「d4」という識別子が付与されている。また、「id="h21"」の見出し要素(head)と「id="h22"」の見出し要素(head)との内容に「SGML」のキーワードが含まれている。

【００８１】図４に示した文書３１に加え、図１７，図１８に示した文書７１，８１に対して関連付け処理が行われると、図８に示した文書５１は以下のように、他の文書の表題もしくは見出しに関連付けられる。

【００８２】図１９は、キーワードと表題もしくは見出しとの関連付けを行った文書の例を示す図である。この図に示すように、文書５４は、他の複数の文書の表題もしくは見出しに関連付けられている。すなわち、図１９において、キーワード「SGML」に対してそれをタグ付けするlink要素の属性によって、文書「d1」の表題「t1」（内容は「SGMLによる電子出版」）、文書「d3」の表題「t1」（内容は「SGMLへの招待」）、見出し「h2」（内容は「SGMLとHTML」）および見出し「h3」（内容は「SGMLとXML」）、文書「d4」の見出し「h21」（内容は「SGML文書の検索」）および見出し「h22」（内容は「SGMLデータベースシステム」）の合計６個の表題もしくは見出しが関連付けられている。

【００８３】以下、このように関連付けられている文書群を対象として、図１６に示したフローチャートに沿って関連付け参照の処理の流れを説明する。まず利用者が文書蓄積部１１に蓄積された文書群から抽出する文書を入力部２１により指示すると、文書抽出部１６は、指示された文書を抽出し、表示部２０に表示する（ステップＳ２１）。ここで表示部２０に表示される文書は図１９に示した文書５４であるものとする。図１９に示す文書５４を表示部２０に表示した場合、link要素の属性値は画面上に表示されないので、第１の実施の形態の場合と同じく図１２に示すように表示画面６１が表示される。

【００８４】次に、利用者が表示部２０に表示された文書５４を参照し、入力部２１より関連付けの付与されたキーワード「SGML」の表示箇所をマウスでクリックするなどの方法で選択する（ステップＳ２２）。見出し抽出部１７は、ステップＳ２２にて選択されたキーワードの関連付け情報すなわちlink要素の属性「ref」の識別子を参照し、文書蓄積部１１から該当する識別子を持つ文書の表題もしくは見出しを抽出する（ステップＳ２３）。

【００８５】次に、見出し抽出部１７は、ステップＳ２３にて抽出された表題もしくは見出しが１つであるか複数であるかを判定する（ステップＳ２４）。図１９に示した例では、合計６個の表題もしくは見出しが抽出されるので、ステップＳ２５へ進む。

【００８６】次に、見出し抽出部１７は、ステップＳ２４にて抽出された表題もしくは見出しが複数あると判定されると、それらの表題もしくは見出しを文書ごとにグループ化する（ステップＳ２５）。図１９の文書５４では、文書「d1」の表題「t1」を１つのグループに、文書「d2」の表題「t1」、見出し「h2」および見出し「h3」を１つのグループに、文書「d3」の見出し「h21」および見出し「h22」を１つのグループにまとめる。

【００８７】このように、抽出された表題もしくは見出しを文書ごとにグループ化することで、同一文書内の関連する記述を連続して参照することができるようになる。次に、見出し抽出部１７は、ステップＳ２５にてまとめられた文書ごとの関連付けのグループを、同一文書内への関連付けの数、および関連付けられる表題もしくは見出しの階層の深さから算出される重要度に応じて並べ替える（ステップＳ２６）。本実施の形態では文書ごとの重要度を次の式によって算出する。

【００８８】

【数１】

$$\text{重要度} = \sum_{i=1}^{n} 2^{-d_i} \quad \cdots\cdots (1)$$

【００８９】式（１）において、ｎは、その文書で関連付けられている表題もしくは見出しに対して１から順に割り振られた数字の最大値を表す。ｄｉは、数字（ｉ）が割り振られた表題もしくは見出しの階層構造における深さを表す（表題の深さを０とする）。すなわち、表題についてはｄｉ＝０、第１レベルの見出しについてはｄｉ＝１、第２レベルの見出しについてはｄｉ＝２などとなる。式（１）に従って各文書の重要度を計算すると、図６に示した文書３３は表題「t1」が１つだけ関連付けられているので重要度＝2^{-0}＝１、図１７に示した文書７１は表題「t1」、見出し「h2」および見出し「h3」の３つが関連付けられているので重要度＝$2^{-0}+2^{-1}+2^{-1}$＝２、図１８に示した文書８１は見出し「h21」および見出し「h22」の２つが関連付けられているので重要度＝$2^{-2}+2^{-2}$＝０．５となる。したがって、文書ごとの重要度にしたがって文書「d2」、文書「d1」、文書「d3」の順に関連付けのグループを並べ替える。

【００９０】なお、文書ごとの重要度の算出方法は、式（１）に示したものに限定されるわけではない。関連付けられる表題もしくは見出しが多いほうが重要度がより高くなるように、また、関連付けられる表題もしくは見出しの階層の深さが浅いほうが重要度がより高くなるように重要度を決めればよい。このような重要度の決定方法は、同一文書内で関連付けられる表題もしくは見出しが多いほうが、そのキーワードが文書全体の主題に関係する可能性が高いと考えられ、また、関連付けられる表題もしくは見出しの階層の深さが浅いほうが、そのキーワードについてより包括的に説明されている可能性が高いと考えられるので、有効な方法である。

【００９１】次に、見出し抽出部１７は、ステップＳ２５にて文書ごとにグループ化された関連付けを、関連付けられる表題もしくは見出しの階層の深さから算出される重要度に応じて各グループ内で並び替える（ステップＳ２７）。本実施の形態では、階層の深さが浅いほうが重要度が高いものとする。また、階層の深さが同一である場合には、文書中で先に出現するほうが重要度が高いものとする。あるいは、文書中での出現順序を優先した重要度を用いてもよい。

【００９２】以上の処理が行われた後、抽出された表題もしくは見出しが表示部２０に表示される。図２０は、複数の見出しを表示する表示画面の例を示す図である。これは、図１２に示した表示画面６１中でキーワード「SGML」を選択したときに表示される表示画面１０１の例を示したものである。図２０に表示されている表題もしくは見出しは、上記処理により、文書ごとにグループ化され、重要度順に並べ替えられている。

【００９３】次に、利用者は表示部２０に表示されている複数の表題もしくは見出しから入力部２１により１つを選択する（ステップＳ２８）。すると、見出し抽出部１７は、ステップＳ２３にて抽出された表題もしくは見出しが１つである場合またはステップＳ２８にて見出しが選択された場合に、その表題もしくは見出しに関連付けられた下位の見出しが存在するかどうかを判定する（ステップＳ２９）。

【００９４】このように、関連付けられる表題もしくは見出しが同一文書内に複数存在する場合や、関連付けられる表題もしくは見出しを持つ文書が複数存在する場合に、重要なものから優先的に参照できるので、たとえ１つのキーワードに多量の文書の表題や見出しが関連付けられている場合でも、効率的に関連付けられた内容を参照することができる。

【００９５】なお、上記の処理機能は、コンピュータによって実現することができる。その場合、文書関連付け装置及び文書閲覧装置が有すべき機能の処理内容は、コンピュータで読み取り可能な記録媒体に記録されたプログラムに記述しておく。そして、このプログラムをコンピュータで実行することにより、上記処理がコンピュータで実現される。コンピュータで読み取り可能な記録媒体としては、磁気記録装置や半導体メモリ等がある。市場に流通させる場合には、ＣＤ－ＲＯＭ(Compact Disk Read Only Memory) やフロッピーディスク等の可搬型記録媒体にプログラムを格納して流通させたり、ネットワークを介して接続されたコンピュータの記憶装置に格納しておき、ネットワークを通じて他のコンピュータに転送することもできる。コンピュータで実行する際には、コンピュータ内のハードディスク装置等にプログラムを格納しておき、メインメモリにロードして実行する。

【００９６】

【発明の効果】以上説明したように、本発明の文書関連付け装置では、文書中のキーワードと被関連付け対象文書の処理対象要素とを関連付けるとともに、被関連付け対象文書中の要素の上位構造と下位構造とを関連付けるようにしたため、文書中のキーワードから他の文書中の要素及びその要素の下位構造を順次辿ることができ、必要最小限の関連付けられた内容を参照することができる。しかも、特定の要素からのみキーワードの抽出を行うため、キーワード抽出に伴う複雑な処理を限られた範囲に対して実行することができ、関連付け処理を高速に行うことが可能となる。

【００９７】また、本発明の文書閲覧装置では、文書中のキーワードと被関連付け対象文書の処理対象要素とを関連付けるとともに、被関連付け対象文書中の要素の上位構造と下位構造とを関連付けておき、文書中のキーワードが指定されると、そのキーワードの関連要素の内容とその下位構造の内容を抽出するようにしたため、キーワードを指定したユーザは、そのキーワードに関する必要最小限の関連要素の内容を参照することができる。

【００９８】また、本発明の文書関連付けプログラムを記録したコンピュータ読み取り可能な記録媒体では、記録された文書関連付けプログラムをコンピュータに実行させることにより、文書中のキーワードと被関連付け対象文書の処理対象要素とを関連付けるとともに、被関連付け対象文書中の要素の上位構造と下位構造とを関連付ける処理を、コンピュータに高速に行わせることが可能となる。すなわち、文書中のキーワードを他の文書の最小限の関連記述に関連付ける処理を、コンピュータに高速に行わせることができる。

【００９９】また、本発明の文書閲覧プログラムを記録したコンピュータ読み取り可能な記録媒体では、記録された文書閲覧プログラムをコンピュータに実行させることにより、文書中のキーワードと被関連付け対象文書の処理対象要素とを関連付けるとともに、被関連付け対象文書中の要素の上位構造と下位構造とを関連付けておき、文書中のキーワードが指定されると、そのキーワードの関連要素の内容とその下位構造の内容を抽出するような処理をコンピュータに行わせることが可能となる。すなわち、コンピュータに対してキーワードを指定したユーザは、そのキーワードに関する必要最小限の関連要素の内容を参照することができる。

【図面の簡単な説明】

【図１】　本発明の原理構成図である。

【図２】　本発明を適用した文書閲覧装置の構成を示す図である。

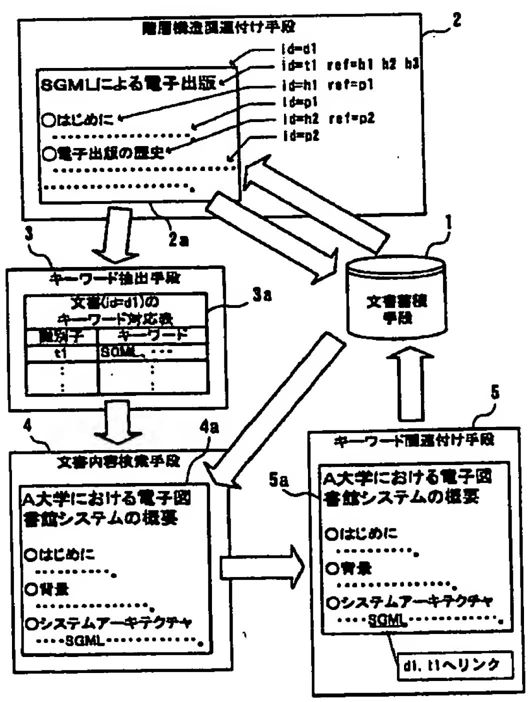
【図３】　文書間の関連付けを行う手順を示すフローチャートである。

【図４】　関連付けの対象となるキーワードを見出しに含む文書の第１の例を示す図である。

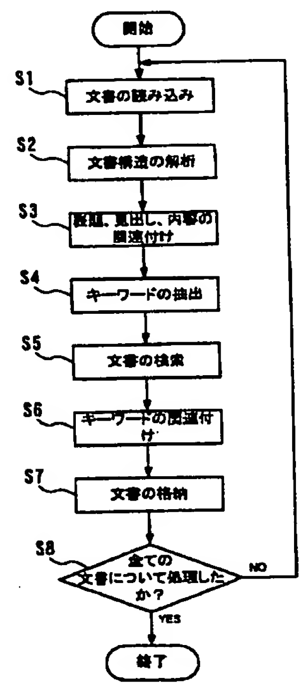
【図５】　各要素に一意な識別子を付与した文書を示す図である。

【図６】　表題、見出し、内容を関連付けた文書の例を示す図である。
【図７】　キーワード対応表の例を示す図である。
【図８】　関連付けの対象となるキーワードを本文中に含む文書の例を示す図である。
【図９】　キーワードと表題との関連付けが行われた文書の例を示す図である。
【図１０】　図９の文書に対して階層構造の関連付けを行った結果を示す図である。
【図１１】　関連付けの利用手順を示すフローチャートである。
【図１２】　文書の内容を表示した際の表示画面の例を示す図である。
【図１３】　見出しを表示した際の表示画面の例を示す図である。
【図１４】　下位の見出しを表示した際の表示画面の例を示す図である。
【図１５】　内容を表示した際の表示画面の例を示す図である。
【図１６】　第２の実施の形態における関連付け参照の処理の流れを示すフローチャートである。
【図１７】　関連付けの対象となるキーワードを表題に含む文書の第２の例を示す図である。
【図１８】　関連付けの対象となるキーワードを表題に含む文書の第３の例を示す図である。
【図１９】　キーワードと表題もしくは見出しとの関連付けを行った文書の例を示す図である。
【図２０】　複数の見出しを表示する表示画面の例を示す図である。
【符号の説明】
１　文書蓄積手段
２　階層構造関連付け手段
２ａ　被関連付け対象文書
３　キーワード抽出手段
３ａ　キーワード対応表
４　文書内容検索手段
４ａ　文書
５　キーワード関連付け手段
５ａ　文書

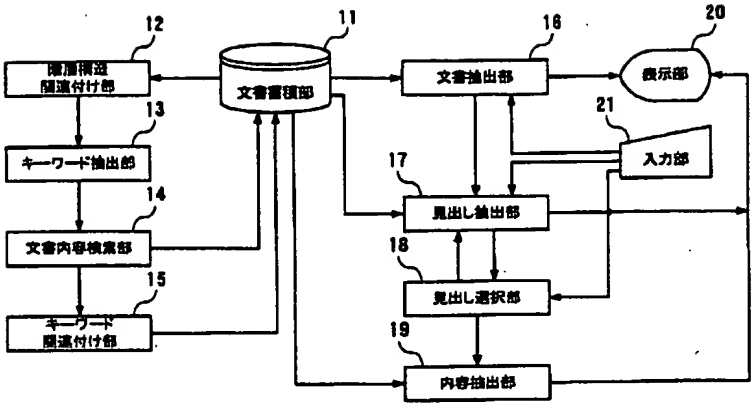
【図１】



【図３】



【図２】



【図４】

31

```
<doc>
  <title>SGMLによる電子出版</title>
  <sect1>
    <head>はじめに</head>
    <para>
      ...
    </para>
    ...
  </sect1>
  <sect1>
    <head>電子出版の歴史</head>
    <para>
      ...
    </para>
  </sect1>
  <sect1>
    <head>関連ツール</head>
    <para>ここでは、SGMLによる電子出版を実現する際に利用できる
    ツールについて説明する。</para>
    <para>...</para>
    ...
  </sect1>
  ...
</doc>
```

【図７】

41

要素の種類	識別子	キーワード
表題	t1	SGML、電子、出版
見出し	h1	－
見出し	h2	電子、出版、歴史
見出し	h3	関連、ツール

【図８】

51

```
<doc>
  <title>A大学における電子図書館システムの概要</title>
  <sect1>
    <head>はじめに</head>
    <sect2>
      <head>背景</head>
      <para>
        ...
      </para>
    </sect2>
    <sect2>
      <head>目的</head>
      <para>
        ...
      </para>
    </sect2>
  </sect1>
  <sect1>
    <head>システムアーキテクチャ</head>
    <para>
    ...
    </para>
    <para>
    ...SGMLへ変換する。...
    </para>
  </sect1>
  ...
</doc>
```

【図５】

32

```
<doc id="d1">
  <title id="t1">SGMLによる電子出版</title>
  <sect1 id="s1">
    <head id="h1">はじめに</head>
    <para id="p1">
      ...
    </para>
    ...
  </sect1>
  <sect1 id="s2">
    <head id="h2">電子出版の歴史</head>
    <para id="p2">
      ...
    </para>
  </sect1>
  <sect1 id="s3">
    <head id="h3">関連ツール</head>
    <para id="p3">ここでは、SGMLによる電子出版を実現する際に利用できる
    ツールについて説明する。</para>
    <para id="p4">...</para>
    ...
  </sect1>
  ...
</doc>
```

【図６】

33

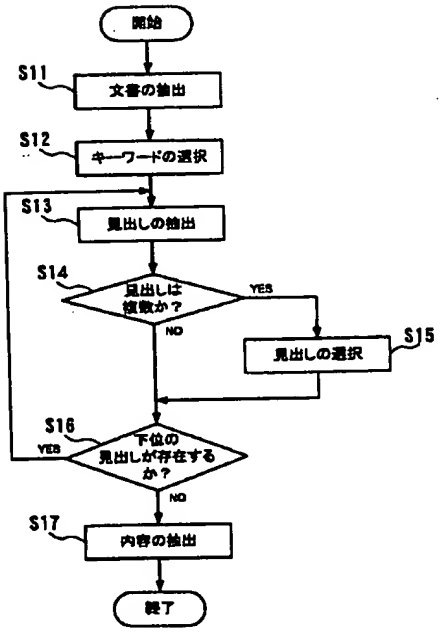
```
<doc id="d1">
  <title id="t1" ref="h1 h2 h3 ...">SGMLによる電子出版</title>
  <sect1 id="s1">
    <head id="h1" ref="p1 ...">はじめに</head>
    <para id="p1">
      ...
    </para>
    ...
  </sect1>
  <sect1 id="s2">
    <head id="h2" ref="p2 ...">電子出版の歴史</head>
    <para id="p2">
      ...
    </para>
  </sect1>
  <sect1 id="s3">
    <head id="h3" ref="p3 p4 ...">関連ツール</head>
    <para id="p3">ここでは、SGMLによる電子出版を実現する際に利用できる
    ツールについて説明する。</para>
    <para id="p4">...</para>
    ...
  </sect1>
  ...
</doc>
```

【図９】

52

```
<doc>
  <title>A大学における電子図書館システムの概要</title>
  <sect1>
    <head>はじめに</head>
    <sect2>
      <head>背景</head>
      <para>
        ...
      </para>
    </sect2>
    <sect2>
      <head>目的</head>
      <para>
        ...
      </para>
    </sect2>
  </sect1>
  <sect1>
    <head>システムアーキテクチャ</head>
    <para>
      ...
    </para>
    <para>
      <link ref="d1.t1">SGML</link>へ変換する。...
    </para>
  </sect1>
  ...
</doc>
```

【図１１】



【図１０】

53

```
<doc id="d2">
  <title id="t1" ref="h1 h2 ...">A大学における電子図書館システムの概要</title>
  <sect1 id="s1">
    <head id="h1" ref="h11 h12">はじめに</head>
    <sect2 id="s11">
      <head id="h11" ref="p1">背景</head>
      <para id="p1">
      ...
      </para>
    </sect2>
    <sect2 id="s12">
      <head id="h12" ref="p2">目的</head>
      <para id="p2">
      ...
      </para>
    </sect2>
  </sect1>
  <sect1 id="s2">
    <head id="h2" ref="p3 p4">システムアーキテクチャ</head>
    <para id="p3">
    ...
    </para>
    <para id="p4">
    ...<link ref="d1.t1">SGML</link>へ変換する。...
    </para>
  </sect1>
  ...
</doc>
```

【図１２】

61

A大学における電子図書館システムの概要

1. はじめに
1.1 背景
...
1.2 目的
...
2. システムアーキテクチャ
...
... SGML へ変換する。...
...

【図１３】

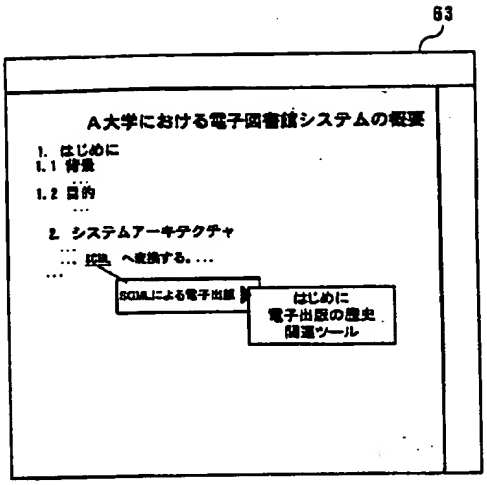
62

A大学における電子図書館システムの概要

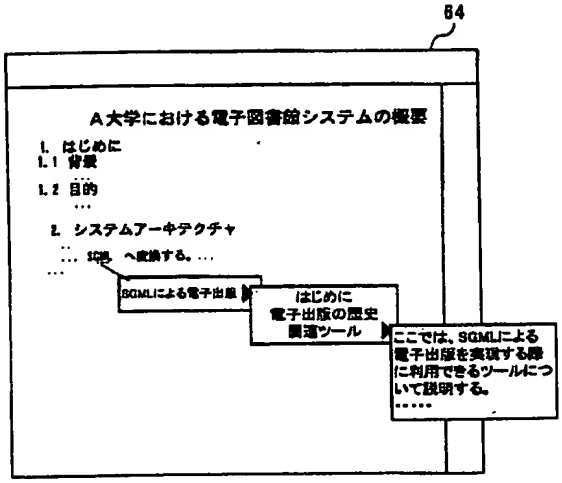
1. はじめに
1.1 背景
...
1.2 目的
...
2. システムアーキテクチャ
...
... SGML へ変換する。...
...

SGMLによる電子出版

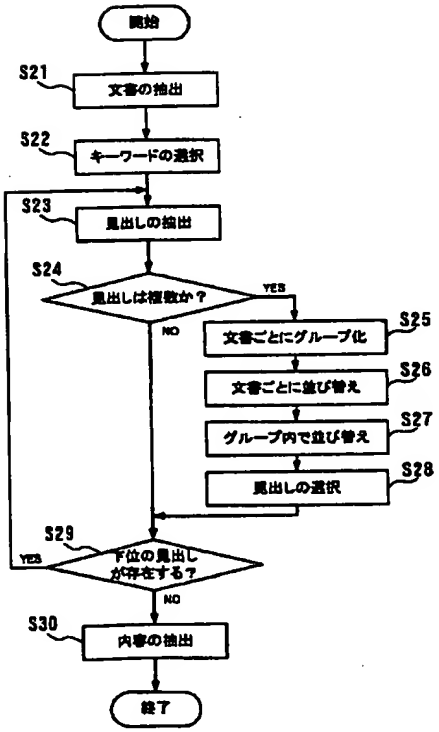
【図１４】



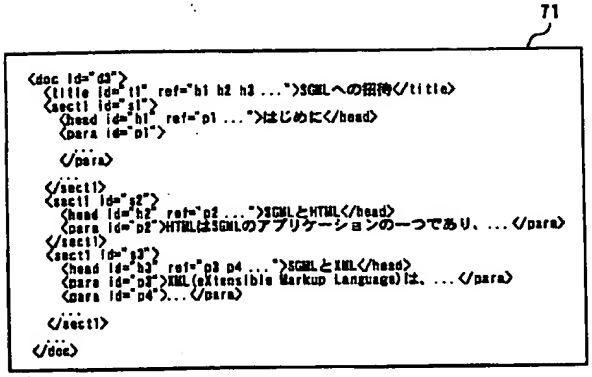
【図１５】



【図１６】



【図１７】



【図１８】

81

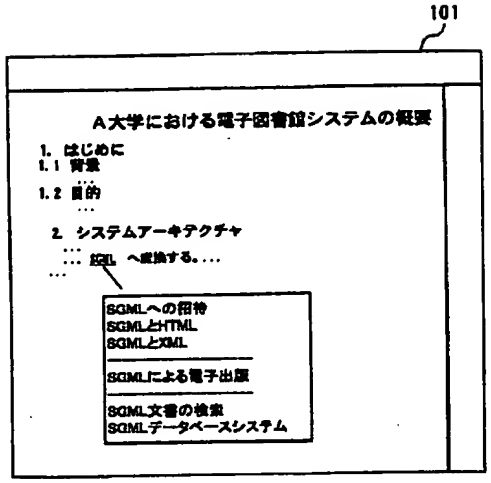
```
<doc id="d4">
  <title id="t1" ref="h1 h2 h3 ...">情報検索技術の最新動向</title>
  <sect1 id="s1">
    <head id="h1" ref="p1 ...">はじめに</head>
    <para id="p1">
      ...
    </para>
    ...
  </sect1>
  <sect1 id="s2">
    <head id="h2" ref="h21 h22">構造化文書の検索</head>
    <sect2 id="s21">
      <head id="h21" ref="p3 ...">SGML文書の検索</head>
      <para id="p3">...</para>
      ...
    </sect2>
    <sect2 id="s22">
      <head id="h22" ref="p9 ...">SGMLデータベースシステム</head>
      <para id="p9">...</para>
      ...
    </sect2>
  </sect1>
  ...
</doc>
```

【図１９】

54

```
<doc id="d2">
  <title id="t1" ref="h1 h2 ...">A大学における電子図書館システムの概要</title>
  <sect1 id="s1">
    <head id="h1" ref="h11 h12">はじめに</head>
    <sect2 id="s11">
      <head id="h11" ref="p1">背景</head>
      <para id="p1">
        ...
      </para>
    </sect2>
    <sect2 id="s12">
      <head id="h12" ref="p2">目的</head>
      <para id="p2">
        ...
      </para>
    </sect2>
  </sect1>
  <sect1 id="s2">
    <head id="h2" ref="p3 p4">システムアーキテクチャ</head>
    <para id="p3">
      ...
    </para>
    <para id="p4">
      <link ref="d1.t1 d2.t1 d3.h2 d3.h3 d4.h21 d4.h22">SGML</link>へ変換する。...
    </para>
  </sect1>
  ...
</doc>
```

【図２０】


101
A大学における電子図書館システムの概要
1. はじめに
1.1 背景
1.2 目的
2. システムアーキテクチャ
SGML へ変換する....
SGMLへの招待
SGMLとHTML
SGMLとXML
SGMLによる電子出版
SGML文書の検索
SGMLデータベースシステム